Labelling the Behaviour of Local Descriptors for Selective Video Content Retrieval

Julien Law-To, Valérie Gouet-Brunet, Olivier Buisson, Nozha Boujemaa

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Labelling the Behaviour of Local Descriptors for Selective Video Content Retrieval

Julien Law-To — Valerie Gouet-Brunet — Olivier Buisson — Nozha Boujemaa

N° 5821

January 2006

Thème COG

ISSN 0249-6399 ISRN INRIA/RR--5821--FR+ENG

Labelling the Behaviour of Local Descriptors for Selective Video Content Retrieval

Julien Law-To[*†], Valerie Gouet-Brunet[†] , Olivier Buisson[*] , Nozha Boujemaa[†]

Thème COG —Systèmes cognitifs
Projet Imedia

Rapport de recherche n° 5821 —January 2006 — 22 pages

Abstract: This paper presents an approach for indexing a large set of videos by considering the cinematic behaviour of local visual features along the sequences. The proposed concept is based on the extraction and the local description of interest points and further on the estimation of their trajectories along the video sequence. Analysing the low-level description obtained allows to highlight semantic trends of behaviours and then to assign labels. Such an indexing approach of the video content has several interesting properties: the low-level description provides a rich and compact description, while labels of behaviour provide a generic and semantic description, relevant for selective video content retrieval depending on the application. The approach is firstly evaluated for Content-Based Copy Detection. We show that taking these labels into account allows to significantly reduce false alarms. Secondly, the approach is experimented on particular applications of video monitoring, where selective labels of behaviour show their capability to improve the analysis and the retrieval of spatio-temporal video content.

Key-words: Video indexing, content based copy detection, local descriptors, trajectories of local descriptors, cinematic behaviours, semantic labels.

[*] INA
[†] IMEDIA Research Group-INRIA

Etiquetage du comportement de descripteurs locaux pour une recherche sélective de contenus vidéo

Résumé : Ce document présente une approche pour l'indexation d'un grand nombre de vidéos, en tenant compte du comportement cinématique de caractéristiques visuelles locales extraites des images de la vidéo. Le concept proposé est basé sur l'extraction et la description locale des points d'intérêt permettant l'estimation de leur trajectoire le long de la séquence vidéo. L'analyse de la description de bas niveau obtenue fait ressortir des tendances de comportements, qui permettent d'assigner un label de comportement à chaque descripteur local. Une telle approche d'indexation a plusieurs propriétés intéressantes: d'une part, la description de bas niveau fournit une description riche, compacte et générique du contenu vidéo, et d'autre part les labels de comportement en fournissent une description haut niveau plus sémantique, permettant une recherche sélective de contenu vidéo pertinente pour l'application considérée. L'approche est tout d'abord évaluée pour la détection de copies dans les contenus audiovisuels. Nous montrons que l'utilisation de labels permet de réduire de manière significative le taux de fausses alarmes. Ensuite, l'approche est expérimentée sur des applications particulières de monitorage télévisuel. Nous montrons que le choix de labels pertinents apporte une amélioration notable dans l'analyse et la recherche de contenus audiovisuels, tout en assurant un certaine compacité en terme de nombre de descripteurs nécessaires.

Mots-clés : Indexation vidéo, détection de copies vidéos, descripteurs locaux, trajectoire de descripteurs locaux, labels sémantiques.

1 Introduction

In this paper, we focus on Content-Based Copy Detection (CBCD) on large collection of videos (several hundred hours of videos). It is an application of Content-Based Image Retrieval (CBIR) in vogue, due to the increasing broadcasting of multimedia contents. It is an alternative to the watermarking approach for persistent identification of images and video clips. As opposed to watermarking, CBCD uses a content-based comparison between the original object and the candidate one [6, 10]. It generally consists in extracting few small pertinent features (called signatures or fingerprints) from the image or the video stream and matching them with a database. Several kinds of techniques have been proposed in the literature for the video retrieval: [9] uses a temporal fingerprints based on the cuts in a video sequence whereas [6] compares global descriptions of the video (motion, color and spatio temporal distribution of intensities). For still image retrieval, [2] defines fingerprints based on the wavelets to find replicate images on the web whereas [11] uses local description on *points of interest*. Initially proposed for stereovision purposes, points of interest are sites in an image where the signal takes high frequency in several directions. Using such primitives is mainly motivated by observing that they provide the most compact representation of the image content by limiting the correlation and redundancy between the detected features. When considering image transformations like geometric changes (cropping or shifting), signatures based on *points of interest* have been proved to be efficient for retrieving still images [1] and video sequences [10]. We will revisit them in section 2.

Towards a semantic description of the video contents

The concept we propose involves the estimation and characterization of trajectories of points of interest along the video sequence. Building trajectories of points in videos is a recent topic for video content indexing. At present, such trajectories are usually analysed for modelizing the variability of points along the video and then enhancing their robustness, for generic object recognition (see for example [5, 24]) or for more specific object recognition [23] or also for segmenting and modelling moving objects [19]. In this work, we plan on taking advantage of such trajectories for indexing the *behaviour* of points of interest. First, the trajectory properties will allow to enrich the local description with a spatial, cinematic and temporal behaviour of this point ; second, the redundancy of the local description along the trajectory can be efficiently summarized without loss of information. Analysing the trajectories obtained allows to highlight trends of behaviours and then to assign a label of behaviour to a local descriptor, in function of the application. The aim is to provide a *rich*, *compact* and *generic* video content description, that allows to define *specific* labels for selective retrieval according to the application needs. Indexing the video sequence will consist in producing a mid-level description of the video by extracting points of interest, characterizing them with a local description and a trajectory. This description will be used to define labels of point behaviours, and then will provide a high level description of the video sequence, aiming at a semantic description for the desired application.

Figure 1 illustrates the concept. The different trends of behaviours of local descriptors that can be observed in this sequence reveal different visual and semantic components of the image (the speaker, its attitude, the background and the frame insert). The labels of behaviour that can be defined allow

to exhibit a high level interpretation of the sequence. We feel that exploiting them while indexing the visual contents of the video sequence will allow to improve CBCD and also to define specific queries for selective spatio-temporal video retrieval.

Other approaches exist and involve spatio-temporal points of interest. These specific points are used for global behaviour recognition in a similar way that usual points of interest are used for object recognition. For example, I. Laptev and T. Lindeberg in [12] have described a way of finding spatio-temporal points of interest for motion interpretation. In [4], the authors use a similar approach in order to separate behaviours such as facial expression or mouse activity. Our goal is different in the sense that we use the local behaviour of classical points of interest in order to find similar video sequences or copies.

The paper is structured as follows: in section 2, we present our method to obtain the low-level description of the video sequences, i.e. to extract, to characterize points of interest and to estimate their trajectories along the sequences. Section 3 defines the concept of labels for exhibiting a high-level description of the video, based on the low-level description obtained. Results of an extensive evaluation of the approach applied to CBCD is then described and analysed in section 4. Finally, in section 5 we show the contribution of our approach for video monitoring and more generally for video indexing.

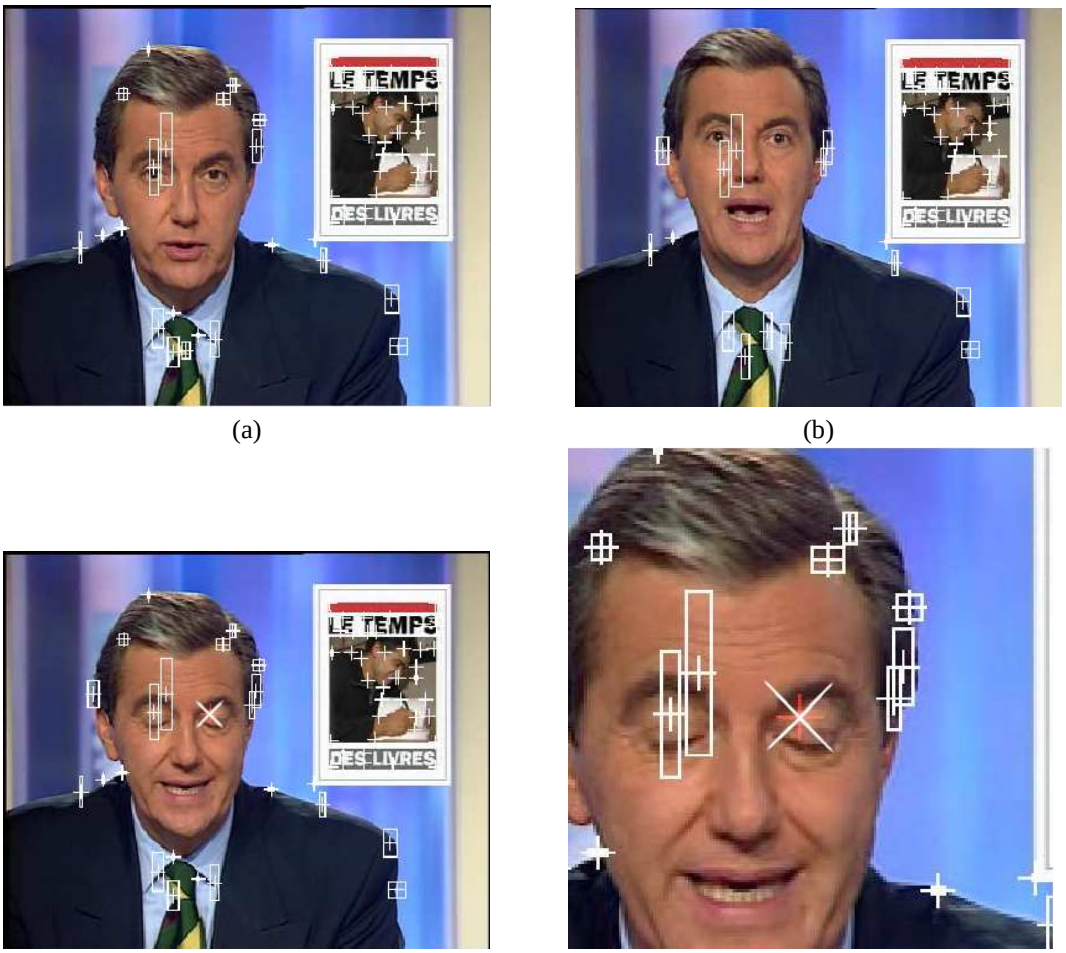
2 Building trajectories of local descriptors

We present here the low-level description of the video sequence. Section 2.1 details the choice we made for interest point extraction and local characterization, while section 2.2 describes the algorithm for tracking these points.

2.1 Extracting and characterizing points of interest

The interesting properties of points of interest make them popular in the literature of Computer Vision and CBIR. The well-known Harris and Stephens detector [7] used to be described with local features, applied to grey value or color images. Many works have been done to make them robust to several image transformations. They have been applied to grey level images [21] and to color images [17]. Then they have been adapted to changes of scale [13, 14] and generalized to affine transformations [15]. A recent performance evaluation [16] has shown that the SIFT descriptor [14] performs best for object recognition. More recently, points of interest have been extended to spatio-temporal signal [12, 4]. Points of interest are relevant for precise retrieval in images, like objects or details. Associated to an adequate voting function, they are robust to occlusion and consequently are interesting for copy detection purposes where several geometric transformations of the image can occur, like cropping or shifting.

We have not used the SIFT descriptor, first because it involves a high dimensional features set (128 items for each key point), making it incompatible with several hundred hours of videos (one hour represents 90000 pictures, involving roughly 3×10^6 local descriptors). Second, this descriptor is invariant to several image transformations, making it efficient for object recognition but not optimal for tracking where consecutive frames differ by small transformations. We have not used



(a) (b)

(c) (d) Screenshot (c)cropped and zoomed

Figure 1: An example of categories of points of interest according to their behaviour: the boxes represent the amplitude of moving points along the trajectory (motionless points do not have box). The "+" correspond to the mean position of such points. The "x" shows a spatio-temporal rare point (the eye blinking).

a spatio-temporal local descriptor because the temporal part would not be relevant for the tracking step describe bellow. Moreover those spatio-temporal points of interest cannot describe all the kinds of informations of the video: for example, a motionless points is also a relevant information that we would like to keep. Points from the background are not detected with this detector despite their importance in the description of the video. Therefore, the descriptor we employed is the Harris detector associated to a simple local description of the points leading to the following 20 dimensional signatures $\vec{S}$:

$$\vec{S} = \left(\frac{\vec{s_1}}{||\vec{s_1}||}, \frac{\vec{s_2}}{||\vec{s_2}||}, \frac{\vec{s_3}}{||\vec{s_3}||}, \frac{\vec{s_4}}{||\vec{s_4}||} \right)$$

where the $\vec{s_i}$ are 5 dimensional sub-signatures computed at 4 different spatial positions around the interest point. Each $\vec{s_i}$ is a differential decomposition of the grey level signal $I(x, y)$ until order 2:

$$\vec{s_i} = \left(\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}, \frac{\partial^2 I}{\partial x \partial y}, \frac{\partial^2 I}{\partial x^2}, \frac{\partial^2 I}{\partial y^2} \right)$$

Such a description is invariant to image translation and to affine illumination changes.

2.2 Tracking points of interest

Temporal approaches of feature point tracking exist for *point trajectory estimation*. Classically, the encountered techniques involve a function cost defined for three consecutive frames. Different linking strategies are applied to find the correspondences and optimize the trajectories. The most popular approach is probably the Kanade-Lucas-Tomasi (KLT) tracker, proposed in 1981 but fully developed later in [25]. It consists in defining good features by examining the minimum eigenvalue of each 2 by 2 gradient matrix, and in tracking them using a Newton-Raphson method of minimizing the difference between the two windows. Another approach is the one developed by Sethi and Jain in 1987 [22] and called Greedy Exchange algorithm (GE). This algorithm is based on a cost function which penalizes the changes of direction and the magnitude of the speed vector. Salari and Sethi [20] solve the missing and spurious measurements problem of the GE approach by introducing phantom points. In [3], the algorithm "IPAN tracker" described is based on the idea of competing trajectories. The paper also presents a performance evaluation of feature point tracking approaches. More recently, probabilistic and multi-solution tracking methods like particle filters in [8], inspired by the Kalman filter, has been developped to track the non-rigid objects and multiple objects or multiple points.

As we focus on low cost computational techniques, the tracking algorithm we have chosen is basic and does not depend on the local description adopted. A L_2 distance is computed from frame to frame between all the points of interest of the frame and all of those from 15 previous frames and the 15 next frames, as illustrated in figure 2. In most of the tracking method described, a model of trajectory is defined and as we want large degrees of liberty in points behaviour, we have not used those models.

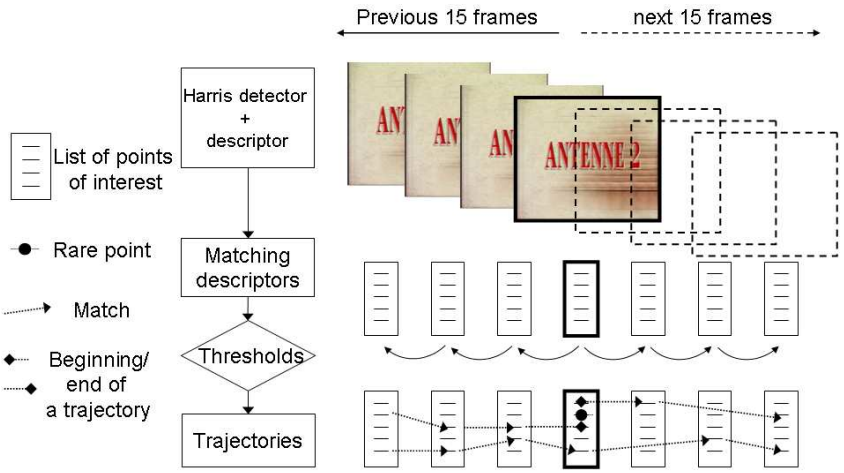


Figure 2: Illustration of the algorithm for trajectory estimation.

A couple of thresholds is used to decide: a matching threshold (T_{match}) and a non matching threshold ($T_{non\ match}$). T_{match} is function of the robustness of the descriptors against usual post-production transformations. $T_{non\ match}$ is obviously higher than T_{match} and allows to produce points called *rare* points. These points do not match but are salient spatially and temporally, making them relevant for video content retrieval. To not mistake them for noise, they are detected with a high threshold of the Harris operator.

For points that match, three decisions can be taken:

- matching only in the future: start of a new trajectory;
- matching only in the past: end of a trajectory;
- matching in the future and in the past: add the point to the trajectory.

3 Labelling local behaviour

In this section, we present and illustrate with an example the choices we have made for building a high-level description of the set of videos, based on the low-level descriptors presented above.

3.1 From low-level to high-level description / From genericity to specialization

3.1.1 Mean local descriptors

At the end of the trajectory building, a *low-level* description of the points of interest based on the signal must be associated to the trajectory. For each trajectory, we take the average of each component of the local descriptors as a low-level description of the trajectory. The descriptor obtained will be noted $\vec{S}_{mean}$ in the rest of the paper, according to section 2.1. As the trajectory is computed from frame to frame, the local signatures may vary along the trajectory. To assess the representativeness of $\vec{S}_{mean}$ in the trajectory, we test on a sequence (see the example of section 3.2) how many local signatures of the trajectory has a distance from $\vec{S}_{mean}$ lower than the matching threshold used during the trajectory building. 95 % of the points of the trajectories have a lower distance from this threshold. This evaluation confirmed that $\vec{S}_{mean}$ is relevant for characterizing a trajectory. A similar approach is described in [5]: the authors show that, on a trajectory, the SIFT descriptor has a quadratic variation depending on the viewing angle, and they take the average of the descriptors in the minimum zone of this variation.

3.1.2 Trajectory parameters

A higher level description of the local descriptors presented above can be obtained by exhibiting the geometric and cinematic behaviour of the interest points. To do this, the following trajectory parameters are stored during the indexing step:

- Average position along the trajectory: μ_x, μ_y;
- Time code of the begin and the end: $[tc_{in}, tc_{out}]$;
- Variation of the position: $[x^{min}, x^{max}]$, $[y^{min}, y^{max}]$.

Added to these characteristics, the mean local descriptors associated to the trajectories provide a richer description of the video content. Such a description is *generic*, because independent of the application. Therefore, it is computed *only once*, whatever the application considered. At this stage, we call the description obtained the *mid-level* description of the video content.

3.1.3 Definition of labels for specialization

From the mid-level description defined above, it is possible to exhibit trends of point behaviours. For example, these categories can be considered:

- Moving points / motionless points;
- Persistent points / rare points;
- Fast motion / low motion points;

- Horizontal motion / vertical motion.

This list is one example of categories of trajectories but we can easily imagine much more others. By classifying the local descriptors according to their behaviour, a label of behaviour can be assigned to them. In the current version of this work, the categories of behaviours are simply obtained by thresholding the parameters defined in section 3.1.2.

Unlike the mid-level description, the labels of behaviour strongly depend on the application considered. According to the other descriptions, it is a *high-level* description, because involving a semantic interpretation of the video, and at the same time it is a *specific* description of the video content, because relevant for selective video content retrieval according to the desired application. This description we obtained is associated with a vote function, that also depends on the application. An example of vote will be described in section 4.1.3 for CBCD.

3.2 An example of descriptors obtained

To give to the reader a primary idea of the low-level to high-level descriptors involved in our approach, we made some statistics on a video sequence: this TV video is composed of commercials and a news show (interviews and reports). Table 1 sums up the statistics obtained. $N_{\vec{S}_{mean}}$ represents the number of mean descriptors according to a specific label. $N_{total_{\vec{S}}}$ represents the total number of points of interest in the video sequence. The label "Motionless" and "Fast Motion" are defined with threshold on the trajectory parameters defined in 3.1.2.

- Label "Motionless" : $(x^{max} - x^{min}) < 5$ and $(y^{min}, y^{max}) < 5)$;
- Label "Fast Motion" : $\frac{(x^{max} - x^{min})}{(Tc_{out} - Tc_{in})} > 8$ or $\frac{(y^{max} - y^{min})}{(Tc_{out} - Tc_{in})} > 8$;

Length of the sequence		1h06min12s
Average length of the trajectories		26
Standard deviation		47
$N_{\vec{S}_{mean}}$ for different labels	total number	per second
All the labels	242 278	61
Rare points	21 447	5.4
Label "Motionless"	130 234	33
Label "Fast Motion"	397	0.1
$N_{total_{\vec{S}}}$	8 229 475	2 072

Table 1: Statistics for 1 hour of video.

The two next sections explain in details the relevance of our approach. In section 4, we will demonstrate that the description proposed in section 3 is richer while more compact than the existing ones for content-based copy detection purposes. Section 5 will show that it can be applied to different applications of video monitoring, where it allows to express specific and selective queries.

4 Evaluation for CBCD

The objective of this section is to demonstrate the relevance of the approach presented in section 3. We particularly focus on the *richness* and *compactness* of the video content description proposed. Here, the context of application is content-based copy detection.

4.1 Framework of the evaluation

4.1.1 The dataset tested

All the experiments are done on 300 hours of videos randomly taken from the video archive database stored at INA (the french *Institut National de l'Audiovisuel*). These videos are TV sequences from several kinds of programs (sports event, news show, talk show) and are stored in *MPEG-1* format with an image size of 352 x 288 pixels.

As a reference we use a symmetrical technique with local description: same algorithm is applied to the database and the queries. This technique uses key frames based on the image activity and local descriptors based on the signal on points of interest. This method presents high performance as shown in [10] even on large video database. We have implemented this technique as a reference rather than [6] because the different globals descriptions (color, motion and distribution of itensities) are not enough robust for our specific needs and so the performance are lower, especially for short video sequences.

4.1.2 Definition of the queries

One video sequence has been randomly chosen from the 300 hours and then modified with classical transformations used in post production like crop, zoom, resize. We also add noise and we change the contrast and the gamma. Figure 3 and tab 2 presents the attacks. Then, we try to find a copy of this attacked video in the whole dataset.

Our retrieval approach is asymmetric so we do not build trajectories with the candidate sequence. The main reason is that the indexing part needs long time computational. So we decide to use as queries Harris points of interest that we choose depending on two parameters:

- period of chosen frame p;
- number of chosen points n per selected frame.

So every p frames the description described in 2.1 is computed on the n points with the highest Harris responses. The position and the time code of the points of interest are also kept in the queries. In order to compare to the symmetrical technique we use $p = 30$ (correspond to the approximatively 0.8 key frame per second of the symmetrical technique) and $n = 20$. The chosen video lasts 24 minutes so it represents 24000 descriptors queries ($\frac{N_{frames}}{p} * n$).

Original Video

Attacked Video

Figure 3: Attacks on the video

zoom	gaussian noise	gamma	contrast	shift
0.96	$\sigma = 5$	$\gamma = 1.1$	1.2	15

Table 2: Parameters of the attacks.

4.1.3 Algorithm of retrieval

By searching the descriptors of the candidate video sequence in the feature space using a statistical similarity search, we measure the similarity between the video sequences. A candidate video clip S can be viewed as a series of K_S points of interest characterized by a spatio temporal position: (tc_m, x_m, y_m) $(m \in [1, K_S])$ with a signal description as defined in 2.1. For each descriptor in range query, the search processing in the reference descriptor index returns a number R_m of results. Each result $r_{m,n}$ $(n \in [1, R_m])$ is a trajectory $([tc^{in}_{m,n}, tc^{out}_{m,n}], [x^{min}_{m,n}, x^{max}_{m,n}], [y^{min}_{m,n}, y^{max}_{m,n}])$ of the descriptor index.

This is a first step to choose some potential matches in the index but it is not discriminant enough so a registration using a geometric model is necessary. The improvement of the registration has been proved (temporal registration [9], spatial registration [11]) and our contribution in the spatio temporal registration is that we have quantified the spatio-temporal liberty of a point with trajectories. When the candidate video clip S is perfectly matching with a R reference clip, there is a spatio-temporal offset. So, during the decision algorithm, this offset is estimated. The main idea is to count the number of queries which are compatible with a given offset. The offset between the query position and the results is then:

$$d_{m,n} = (\left[tc^{in}_{m,n} - tc_m, tc^{out}_{m,n} - tc_m\right], \left[x^{min}_{m,n} - x_m, x^{max}_{m,n} - x_m\right], \left[y^{min}_{m,n} - y_m, y^{max}_{m,n} - y_m\right])$$

As the algorithm is based on interval-valued data, the result is an interval-valued offset and it is compatible with a query if there is at least one result which has an intersection with this offset (O_{ff}):

$$C(O_{ff}, r_m) = \begin{cases} 0 \; if \; \forall n, \; O_{ff} \cap d_{m,n} = \{\emptyset\} \\ 1 \; else \end{cases} \tag{1}$$

This C compatibility measure is applied for each query and the optimal offset maximizes the number of compatible queries:

$$C_r(O_{ff}) = \sum_{m=1}^{K_S} C(O_{ff}, r_m) \tag{2}$$

The C_r criterion is not based on a simple intersection in order to be robust to the outliers. To optimize C_r, the algorithm just tests the different offset $d_{m,n}$ to find the optimal offset. This method is really fast. The search and the vote are computed in less than 1 minute for our tests with 24 000 queries (24 min of video) in our different indexes.

4.2 Richness of the proposed descriptors

The principle of this first experimentation is simple: the video content is indexed by computing a high level description of the video according to the concept presented in section 3.1. For this first experiment, all the local descriptors computed are considered, whatever their labels of behaviour.

The criterion of evaluation is ROC curves that present the recall vs the false alarm rate. The recall for an attacked video sequence depends on a threshold on the number of points found as for the false alarm rate. The false alarm rate is computed by using a video that is not in the tested database and that corresponds to 3 hours of video from a foreign channel. The video used for the false alarm rate represents 180 000 queries.

Our indexing and retrieval technique is then compared to the symmetrical technique using exactly the same parameters as said in 4.1.2. The results are presented in figure 4. We clearly show the improvement for CBCD: with a lower false alarm rate, there is a much better recall rate.

4.3 Richness vs. compactness of the descriptors

The previous experiment clearly shows the improvement of our approach for CBCD facing a state of the art technique, but the size of feature index is higher, with 50.6 descriptors per second of video compared to 16 descriptors per second for the reference technique. It is natural to think that the more we keep descriptors, the best the quality for the recall is. More descriptors cause also more computational times. In order to quantify the size of the feature space, we define the *descriptor rate* r_{desc} as the ratio between the number of mean descriptors in the feature space ($N_{\vec{S}_{mean}}$) according to a specific label and the total number of low-level descriptors ($N_{total_{\vec{S}}}$).

$$r_{desc} = \frac{N_{\vec{S}_{mean}}}{N_{total_{\vec{S}}}} \tag{3}$$

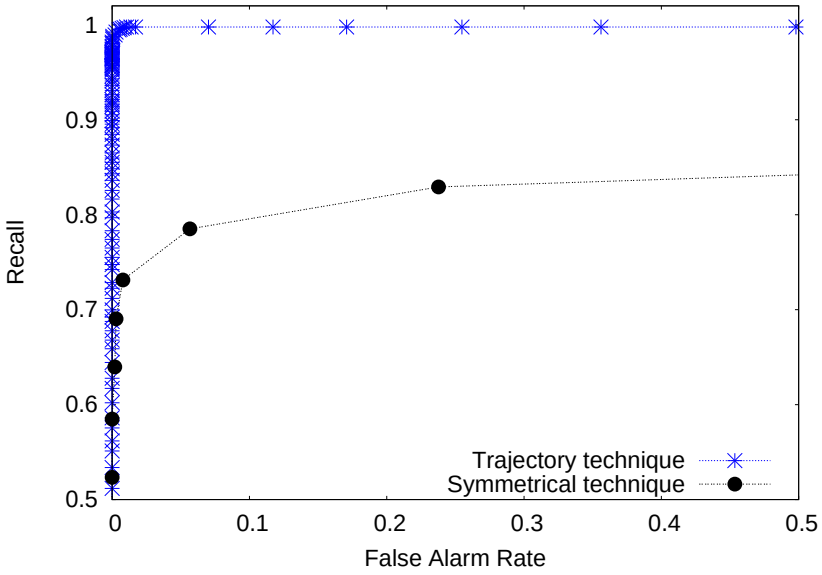


Figure 4: Comparison with the symmetrical technique. All the high-level descriptors are considered, whatever the labels are.

Therefore, particular labels are considered to decrease the size of the feature index and become more selective. The first label considered is the *persistence* of the point along a trajectory because it seems natural to think that the persistence of the points is quite independent of the type of TV program.

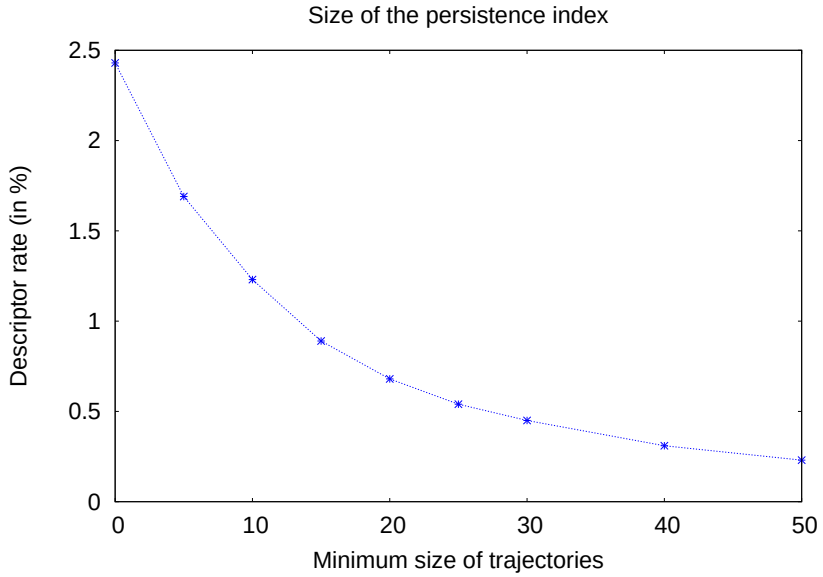


Figure 5: Descriptor rate (r_{desc}) according to labels associated to different persistences.

We measure the influence of the persistence on the video content description by only considering local descriptors labelled as persistent. The different labels are obtained by considering different trajectory lengths. We call $persist_x$ the points with the label $persistence > x$. The figure 5 presents the r_{desc} obtained for several high-level descriptors associated to different persistence.

The ROC curves presented in figure 6 expresses the loss of quality according to the different persistence labels. Several interesting observations can be put down:

- By considering only local descriptors associated to the label "$persist_{20}$", we obtain better results (+ 12 % for the same false alarm rate) than with the symmetrical technique, while the feature space generated has roughly the same size (17.5 millions).
- For trajectories with length lower or equal to 20, results are very similar, while the sizes of the feature spaces involved are very different. For example, for label "$persist_5$", the descriptor rate is 1.5%, while it is 0.70% for label "$persist_{20}$".
- By using very persistent trajectories ("$persist_{50}$" label), the loss of quality is too important and the ROC curve is under the reference one.

This experiment shows the strength of the video description by exploiting trajectories of points of interest. In fact, the use of the tracking to have a full description of each local descriptor enhances the robustness of this description. Sivic in [23] uses a similar method for face recognition in video sequence: by tracking faces along video sequences, he has got a robust and compact description of each face. By using very persistent trajectories, we obtain a very robust description but too specific (background specific to a show, specific to a TV channel ...) and the description is not enough discriminant for a CBCD application.

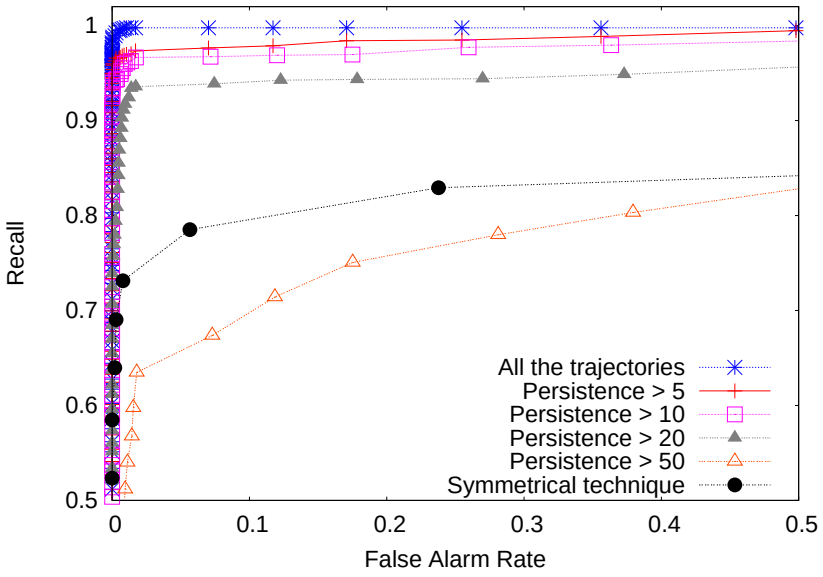


Figure 6: ROC curves for different labels.

4.4 Discussion

These experiments have shown that the proposed approach of video content description is *richer* than the existing ones in the literature. Another main advantage of this approach is that all the steps in the process are very flexible. We can have different strategies for building the high-level descriptors, different strategies for building the queries so we can adapt the system to the needs: for a really precise search with a high granularity, queries would have a very short period for example. The main problem is that building the low-level index costs a lot in computational time. The system is now 3 times slower than the real time with a standard PC (Pentium IV, 2.5 GHz, 1 Go RAM) but we do not have applied computational optimizations on the code. The main CPU costs is the Harris points of interest detection whereas the building trajectories part is very fast. We have to remember

that this mid-level description is only done once and then from this first index it is possible to fastly generate all the high-level descriptors required for the considered application.

5 From genericity to specialization in monitoring applications

In this section, we aim at demonstrating the relevance of our approach for different kinds of queries in monitoring applications.

Determining what is relevant in a video sequence and then what to index clearly depends on the application and moreover may change with time for the same application. The mid-level approach we have proposed indexes the visual content of a video, by caring about being sufficiently generic. Such a description is computed once and for all. The labels of behaviour represent the high-level description of the video contents, they are to be chosen and computed according to a precise context of application. In this section, all the experiments start from the mid-level description of the video content. We discuss the use of these labels for particular applications requiring selective queries.

5.1 Selectivity of the labels of behaviour

The question of knowing what is relevant in a video sequence to fingerprint is still open. In one hand, we can think that the background and the scenes are characteristic for a TV show, but they proved to be not discriminant for CBCD. On the other hand, the local motion of objects in the scene is very discriminant, but not typical of a TV show. In this section, we try to find a way to answer that question by using two kinds of labels based on these observations.

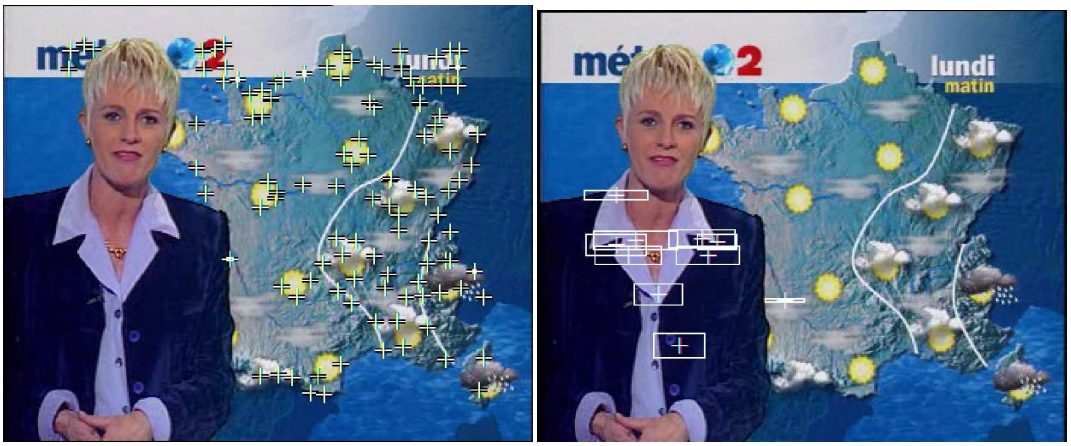
For this experiment, we take a short video sequence (3 min) of a weather forecast program in TV news. This sequence represents a typical problem of false alarm: the background (the map) is common to many sequences of weather forecast programs but these ones cannot be seen as a copy of the others. We use two labels addressing two different categories of behaviour and defined as follows:

- Label 1: Motionless and persistent points
  $(x^{max} - x^{min}) < 5$ and $(y^{min}, y^{max}) < 5)$ and $(Tc_{out} - Tc_{in}) > 35 frames$;
- Label 2: Moving and persistent points
  $((x^{max} - x^{min}) + (y^{max} - y^{min}) > 10)$ and $(Tc_{out} - Tc_{in}) > 15 frames$.

Label 1 is supposed to characterize the background and the TV sets and Label 2 is supposed to describe the motion of moving objects (the speaker in the tested sequence), as illustrated in figure 7.

By using the same framework as in section 4 with the query sequence of figure 7, we obtain interesting results:

- **Query with label 1**. By only exploiting the supposed points from the background, the retrieval algorithm returns 4 different sequences: the first is the original one, the second is a copy of the original (in France, the weather report during the night is the same as in the evening news),



(a) Label 1 (b) Label 2

Figure 7: Query sequence: visualization of the local descriptors and their behaviour according to two different labels. The symbols superimposed must be interpreted as in figure 1.

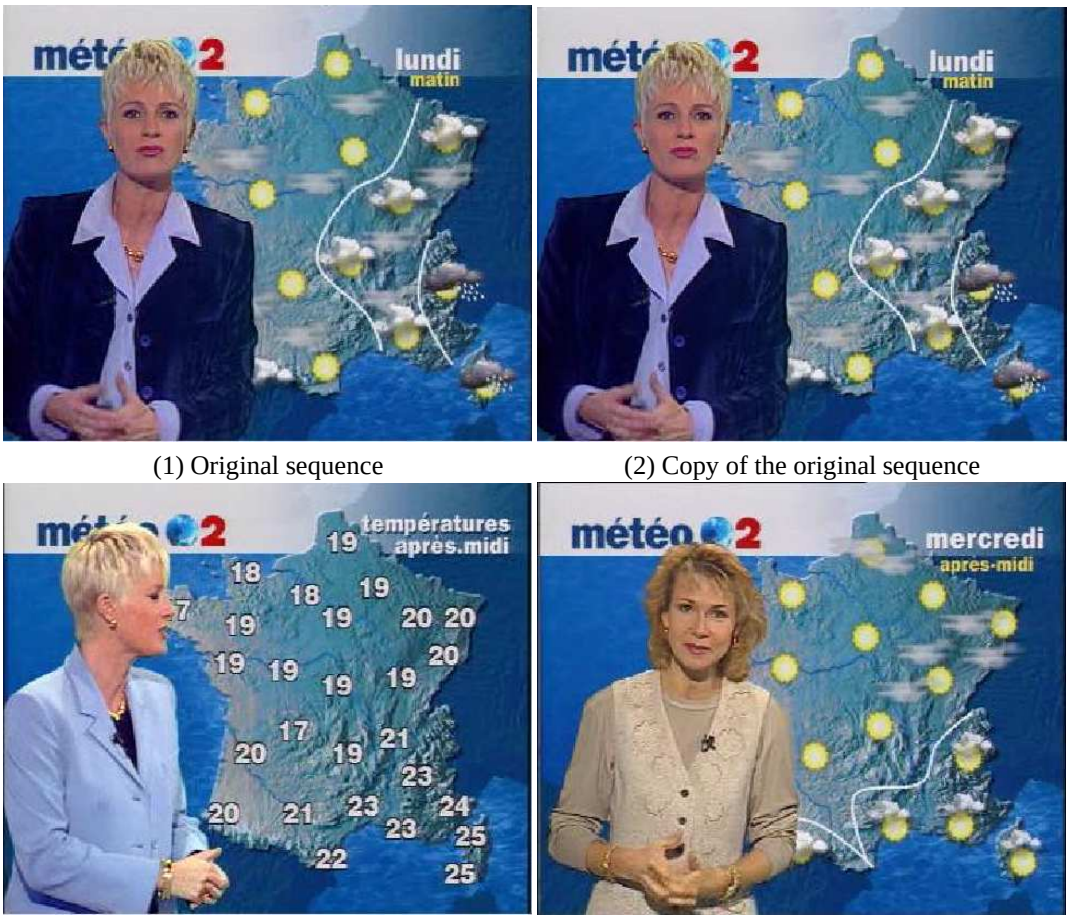
the third is another one weather forecast with the same speaker but with a different costume and the last one is a weather forecast with another speaker (figure 8 presents these results by decreasing order of similarity).

- **Query with label 2**. Exploiting the moving points in the retrieval only returns the original sequence and its copy (the results are only (1) and (2) in figure 8).

This experiment shows that using appropriate labels of behaviour allows to define different levels of monitoring TV programs: we can find TV shows with similarities, and also separate them to keep only the copy. In a CBCD system, retrieving another weather forecast is clearly a false alarm (this was a limit for the symmetrical technique) but in a monitoring application that searches for example, collection of TV programs, it becomes a good detection. Similarly, it is possible to search for soap opera shows, news shows and a lot of recurrent TV programs, just by using a short description of videos with some samples of those programs and selective labels. If the application is to find exactly the same episode of a soap show for example, the use of a label associated to motion proves to be necessary.

5.2 Complementarity of labels of behaviour

Like A. Opelt in [18] who uses different kinds of descriptors together to improve object recognition in still images, the idea is here to exploit jointly different labels to improve CBCD. Using the two previous detailed labels seems relevant because moving and persistent motionless points does not describe the same information. In order to quantify the importance of using different kinds of labels,



(1) Original sequence (2) Copy of the original sequence

(3) False alarm (4) False alarm

Figure 8: Content video retrieval by using selective labels of behaviour.

experiments were conducted by testing four high-level descriptors: label 1, label 2 (from section 5.1), both labels 1 and 2 and also a label "$persist_{38}$" (defined as $persistence > 38$). This last one has been chosen because it has the same descriptor rate as the one of labels 1 + 2 (see table 3).

Labels	r_{desc}	Labels	r_{desc}
Label 1	0.11 %	Label 1+2	0.19 %
Label 2	0.08 %	Label 38	0.20 %

Table 3: Descriptor rates for different labels for 300 hours of video.

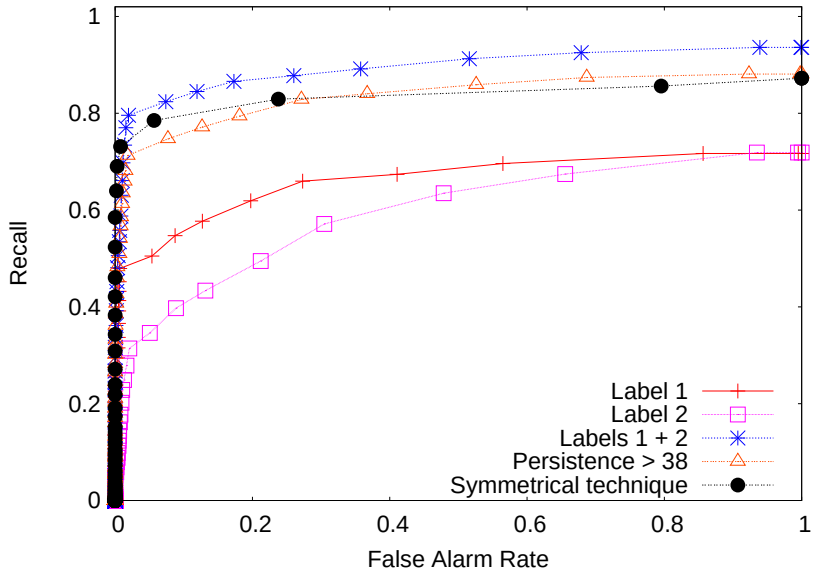


Figure 9: ROC curves for different labels.

The ROC curves presented in figure 9 clearly shows that the joint use of these two labels really improves the retrieval. More precisely:

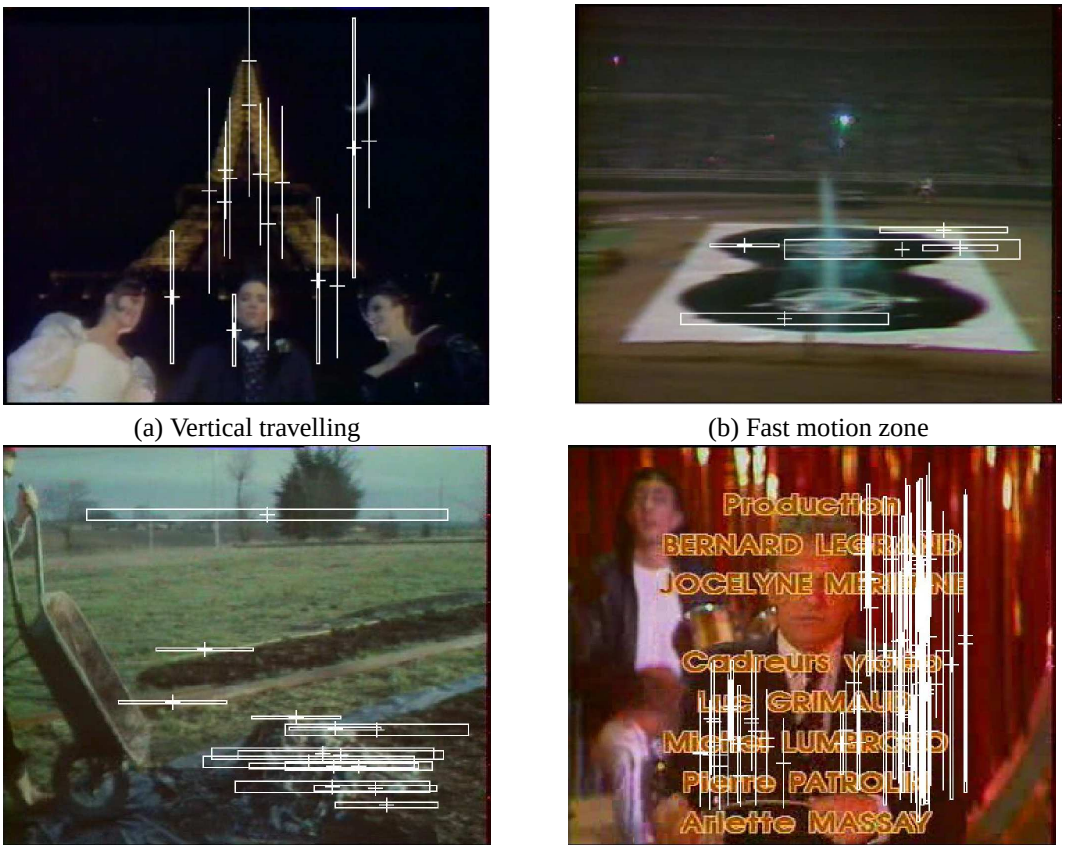
- By using labels 1+2 that generate 7.5 millions of descriptors ($r_{desc} = 0.19\%$), the ROC curve is better than with the symmetrical technique which involves 17.5 millions of descriptors.
- Labels 1+2 are also better than label "$persist_{38}$", while the feature spaces involved have the same size (similar r_{desc}).

In previous sections, we have shown that the selection of an appropriate label of behaviour allows to enhance retrieval for a target application. Here we show that the appropriate combination of several

labels involving different behaviours also enhance retrieval. Such a combination provides a more representative description of what is relevant in the video content while it is more compact.

5.3 Events based segmentation of the video

Another potential application of our approach is an help for the segmentation of the video. Analysing the mid-level description obtained can give useful information of the structure of the video. For example, we can identify spatial events like the credits at the end of the shows (lots of points in the same time codes with vertical motion), travelling, camera motion and also temporal events like cuts (end and start of a lot of trajectories at the same time code). As an illustration, figure 10 shows different video situations that can be highlighted by considering appropriate relevant labels.



(a) Vertical travelling (b) Fast motion zone

(c) Travelling (d) End credits

Figure 10: Video events by analysing the local descriptors behaviours.

6 Conclusion and Perspectives

In this paper, we have presented an approach for selective video content retrieval, based on the smart description of local descriptors behaviours. The experiments performed show that our method presents an interesting potential for different applications such as CBCD. Such an indexing approach has several interesting properties: it is *generic* in the sense that first, it does not suppose any prior knowledge on the video contents ; second as almost all the steps are independent (low-level descriptors computing, tracking and voting), they can be improved, chosen or adapted for the considered application ; and third because the labels of behaviour can be efficiently adapted to the considered application, without having to compute again the mid-level descriptors. Another quality is that the description is *richer* than the classical approaches encountered: using appropriate labels (and appropriate combination of labels) provide a more representative and precise description of the spatio-temporal video content. Such a description contains a high-level interpretation of the video contents, contributing to reduce the semantic gap inherent in visual descriptors. Finally, the approach is more *compact*, as it was demonstrated in this paper.

Future work will consist in an automatic analysis of the mid-level descriptors set, potentially based on non-supervised classification methods. Another direction could be to exploit the correlation between the global motion of the video and the trajectories or even the correlation between trajectories in order to distinguish object motion from camera motion.

References

[1] S.-A. Berrani, L. Amsaleg, and P. Gros. Robust content-based image searches for copyright protection. In *ACM Intl. Workshop on Multimedia Databases*, pages 70–77, 2003.

[2] E. Chang, J. Wang, C. Li, and G. Wilderhold. Rime - a replicated image detector for the world-wide web. In *SPIE Symp. of Voice, Video and data communications*, pages 58–67, 1998.

[3] D. Chetverikov and J. Veresti. Tracking feature points: A new algorithm. In *ICIP*, pages 1436–1438, 1998.

[4] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, October 2005.

[5] M. Grabner and H. Bischof. Extracting object representations from local feature trajectories. In *1st Cognitive Vision Workshop*, 2005.

[6] A. Hampapur and R. Bolle. Comparison of sequence matching techniques for video copy detection. In *Conf. on Storage and Retrieval for Media Databases*, pages 194–201, 2002.

[7] C. Harris and M. Stevens. A combined corner and edge detector. In *4th Alvey Vision Conference*, pages 153–158, 1988.

[8] C. Hue, J. L. Cadre, and P. Perez. Tracking multiple objects with particle filtering. Technical report, IRISA, 2000.

[9] P. Indyk, G. Iyengar, and N. Shivakumar. Finding pirated video sequences on the internet. Technical report, Stanford University, 1999.

[10] A. Joly, C. Frelicot, and O. Buisson. Feature statistical retrieval applied to content-based copy identification. In *ICIP*, 2004.

[11] Y. Ke, R. Sukthankar, and L. Huston. Efficient near-duplicate detection and sub-image retrieval. In *ACM Int. Conf. On Multimedia*, 2004.

[12] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, 2003.

[13] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998.

[14] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157, Corfu, 1999.

[15] K. Mikolajczyk and C. Schmid. An affi ne invariant interest point detector. In *ECCV*, volume 1, pages 128–142, 2002.

[16] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *ICPR*, 2003.

[17] P. Montesinos, V. Gouet, and R. Deriche. Differential Invariants for Color Images. In *Proceedings of* 14^{th} *ICPR*, pages 838–840, Brisbane, Australia, 1998.

[18] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Generic object recognition with boosting, 2004. Technical Report (submitted to [PAMI] 07/04).

[19] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. Segmenting, modeling, and matching video clips containing multiple moving objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC*, June 2004.

[20] V. Salari and I. Sethi. Feature point correspondence in the presence of occlusion. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 12(1):56–73, Jan. 1990.

[21] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. In *Pattern Analysis and Machine Intelligence*, pages 530–535, 1997.

[22] I. K. Sethi and R. Jain. Finding trajectories of feature points in a monocular image sequence. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 9:56–73, 1987.

[23] J. Sivic, M. Everingham, and A. Zisserman. Person spotting: video shot retrieval for face sets. In *Intl. Conf. on Image and Video Retrieval, Singapore*, 2005.

[24] J. Sivic and A. Zisserman. Video data mining using confi gurations of viewpoint invariant regions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC*, June 2004.

[25] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical report CMU-CS-91-132, Apr. 1991.

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifi que
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399